席崇钦,涂冬波,蔡艳. 兼顾能力与知识状态的 Higher-Order CD-CAT 选题方法[J]. 江西师范大学学报(自然科学版),2022,46(2):111-117.
XI Chongqin,TU Dongbo,Cai Yan. The item selection method considering the ability and cognitive profile in higher-order CD-CAT [J]. Journal of Jiangxi Normal University(Natural Science),2022,46(2):111-117.

文章编号:1000-5862(2022)02-0111-07

兼顾能力与知识状态的 Higher-Order CD-CAT 选题方法

席崇钦,涂冬波,蔡　艳*
(江西师范大学心理学院,江西 南昌　330022)

摘要:Higher-order CD-CAT 的选题方法是传统单目标(即只对知识状态自适应)选题方法,这将导致被试能力的测量精度不高. 基于此,在高阶模型和 PWKL 选题方法的框架下,该文开发了适用于 Higher-order CD-CAT 的新选题方法,该方法在选题时能同时兼顾能力和知识状态. 实验结果表明:与传统选题方法相比,新选题方法的能力和知识状态估计精度都更高,并且在题库安全性上也具有明显的优势.

关键词:双目标 CD-CAT;高阶模型;higher-order CD-CAT;选题方法
中图分类号:B 814.7　　**文献标志码**:A　　**DOI**:10.16357/j.cnki.issn1000-5862.2022.02.01

0　引言

计算机化自适应测试(computerized adaptive testing,CAT)是利用现代化信息技术实现的自适应测试形式,通过“量体裁衣”式的自适应技术,满足高效精准的测试需求. 基于项目反应理论的 IRT-CAT,可以高效精准地获取被试的终结性评价(常称为能力),用于评估被试整体表现. 此外,反映被试微观认知结构或认知过程的形成性评价(常称为知识状态或属性掌握模式)也同样重要,这些信息不仅有助于教师更好地“因材施教”[1-2],而且有助于心理学家探索被试在解决问题时的认知过程[3]和精神病学家诊断患者的心理症状[4]. 基于认知诊断理论,研究者们开发出具有认知诊断功能的计算机自适应测验(CD-CAT),以快速精准地获取被试的形成性评价. 另外,在心理与教育的实际应用中,终结性评价和形成性评价同等重要,即同时关注能力和知识状态,使得整体的评价与具体的干预或补救达到和谐统一(如美国“Race to the Top”(RTTT)推行的联邦赠款计划[5]). 为了高效地获取这 2 个方面的信息,研究者们开发出了双目标 CD-CAT(dual-objective CD-CAT)[1,6].

根据采用的模型不同,双目标 CD-CAT 可划分为使用单一模型和分离建模 2 种类型[1,7]. 基于分离建模的双目标 CD-CAT 采用双标定过程(dual-calibration process)标定题库的项目参数,即题库中的每个项目分别由项目反应模型(item response model,IRM)和认知诊断模型(cognitive diagnostic models,CDMs)标定,因此每个项目有 2 套不同的参数. 然而,有研究者[1,8-9]认为应谨慎使用双标定过程标定项目参数,其原因是:IRM 和 CDMs 的潜在结构完全不同,其中 IRM 的潜在结构是连续变量,而 CDMs 的潜在结构是离散变量. 换言之,基于分离建模的双目标 CD-CAT 是一种存在缺陷的、折中的双目标 CD-CAT[1,9]. 为此,C. L. Hsu 等[1]建议双目标 CD-CAT 采用能同时描述能力和知识状态的单一模型(如高阶模型(higher-order cognitive diagnosis models,HO-CDMs))[10],并在此模型基础上提出了 Higher-order CD-CAT. 在 Higher-order CD-CAT 中,被

收稿日期:2022-01-03
基金项目:国家自然科学基金(31960186,31760288,31660278)资助项目.
通信作者:蔡　艳(1979—),女,江西宜春人,教授,博士,博士生导师,主要从事心理统计与测量研究. E-mail:cy1979123@aliyun.com

试的能力由高阶能力表示,该能力主导被试的知识状态,而知识状态又直接影响项目反应. 由于该 CAT 系统的项目参数只由 HO-CDMs 来标定,因此每个项目仅有 1 套参数. 与分离建模相比,单一模型的题库建设成本更低,其原因是:前者的项目要同时拟合 IRM 和 CDMs,条件比较苛刻,而后者的项目只需要拟合 HO-CDMs.

然而,由于尚未开发出适用于 Higher-order CD-CAT 的选题方法,这将导致该 CAT 系统只能采用单目标选题方法(即只对知识状态自适应),这可能导致能力的估计精度不高. 虽然已有对知识状态和能力同时自适应的双目标选题方法[2,6-7,9,11-14],但是这些方法均是建立在分离建模基础上的,无法应用于 Higher-order CD-CAT 中[1,7-8]. 其主要原因是:Higher-order CD-CAT 只标定 1 套项目参数,而当前的双目标选题方法需要 2 套项目参数. 在 Higher-order CD-CAT 中,若需同时精确估计被试的能力和知识状态,则其选题就应同时考虑这 2 个变量. 基于这个思想,本文在后验加权库尔贝克-莱布勒信息量法(Posterior-Weighted Kullback-Leibler,PWKL)选题方法的框架下,结合 HO-CDMs 的原理,提出适用于 Higher-order CD-CAT、兼顾能力和知识状态的新选题方法.

1 认知诊断模型

认知诊断是认知心理学与心理计量学相结合的产物,在 CDMs 中它融合了相关认知变量以实现对被试的诊断与分类[15]. 设向量 $\boldsymbol{X}_i=(X_{ij})$,其中 X_{ij} 为被试 i 在项目 j 上的作答,$i=1,2,\cdots,N$,$j=1,2,\cdots,J$. 设被试 i 的知识状态为 $\boldsymbol{\alpha}_i=(\alpha_{ik})$,其中 $k=1,2,\cdots,K$,当被试 i 掌握属性 k 时,$\alpha_{ik}=1$,否则 $\alpha_{ik}=0$. 项目与属性之间的关系用 $J\times K$ 的 $\boldsymbol{Q}$ 矩阵表示[16];在该矩阵中当第 j 行第 k 列的元素 $q_{jk}=1$ 时,项目 j 测量了属性 k;当 $q_{jk}=0$ 时,项目 j 没有测量属性 k.

拓广 DINA 模型(generalized DINA,G-DINA)为由 de la Torre J[17]在 DINA 模型[18]基础上拓展的饱和模型,该模型同时考虑了属性的主效应和所有可能的交互效应. 目前许多 CDMs 是 G-DINA 模型的简化版本,即通过约束特定条件,使得 G-DINA 可以转换为不同的简化 CDMs. 如当属性仅存在主效应而无交互效应时,G-DINA 可简化为 A-CDM 模型[17].

为了同时描述被试能力 θ 和知识状态 α,de la Torre J 等[10]通过层次框架结构将 θ 和 α 连接起来,开发出了高阶模型(HO-CDMs). 其中被试 θ 主导 α(水平 2),而 α 影响其在项目上的作答结果(水平 1). 若被试的 θ 水平越高则其掌握某个属性的概率越大,能力 θ 和属性 α_k 的关系为

$$P(\alpha_k=1\mid\theta)=\exp(1.7\lambda_{1k}(\theta-\lambda_{0k}))/(1+\exp(1.7\lambda_{1k}(\theta-\lambda_{0k}))),\tag{1}$$

其中 λ_{0k} 为属性 k 的截距,λ_{1k} 为属性 k 的斜率参数,θ 为服从标准正态分布的高阶能力. 在属性间局部独立的假设下,当被试 i 的能力为 θ_i 时,其知识状态为 $\boldsymbol{\alpha}_i$ 的条件概率为

$$P(\boldsymbol{\alpha}_i\mid\theta_i)=\prod_{k=1}^{K}P(\alpha_{ik}\mid\theta_i).\tag{2}$$

在 HO-CDMs 框架下,任何 CDM 都可以作为项目反应函数[10]. 当同时考虑属性的主效应和所有的交互效应时,水平 1 的项目反应函数应采用 G-DINA 模型,此时称之为 HO-GDINA.

2 适用于 Higher-order CD-CAT 的新选题方法

在 HO-CDMs 中,被试 θ 主导 α,而 α 直接影响被试在项目上的作答结果[10]. 由此可知,被试在项目上的作答结果是由 α 和 θ 共同决定的. 因此,Higher-order CD-CAT 的选题不仅对 α 自适应,而且同时对 α 和 θ 自适应,从而兼顾这 2 个变量的估计精度. 当对连续变量 θ 取 R 个节点时,可得到 $R\times 2^K$ 个 α 与 θ 的组合,每一个组合都可能是被试真实的知识状态和能力. 给定被试 i 的已作答向量 $\boldsymbol{X}_i^{t-1}$,基于贝叶斯定理,可以近似计算 $\boldsymbol{\alpha}_c$ 和 θ_r 的联合后验分布,其计算方法为

$$\pi_i^{t-1}(\boldsymbol{\alpha}_c,\theta_r)\propto L(\boldsymbol{X}_i^{t-1}\mid\boldsymbol{\alpha}_c,\theta_r)P(\boldsymbol{\alpha}_c\mid\theta_r)P(\theta_r)/\left(\sum_{r=1}^{R}\sum_{c=1}^{2^K}L(\boldsymbol{X}_i^{t-1}\mid\boldsymbol{\alpha}_c,\theta_r)P(\boldsymbol{\alpha}_c\mid\theta_r)P(\theta_r)\right)=L(\boldsymbol{X}_i^{t-1}\mid\boldsymbol{\alpha}_c)P(\boldsymbol{\alpha}_c\mid\theta_r)P(\theta_r)/\left(\sum_{r=1}^{R}\sum_{c=1}^{2^K}L(\boldsymbol{X}_i^{t-1}\mid\boldsymbol{\alpha}_c)P(\boldsymbol{\alpha}_c\mid\theta_r)P(\theta_r)\right),\tag{3}$$

其中 $P(\theta_r)$ 为 θ_r 的概率密度函数;$P(\boldsymbol{\alpha}_c\mid\theta_r)$ 为在给定 θ_r 的条件下 $\boldsymbol{\alpha}_c$ 的条件概率(见式(2)). 由式(3)可知,似然函数 $L(\boldsymbol{X}_i^{t-1}\mid\boldsymbol{\alpha}_c)$ 与条件概率 $P(\boldsymbol{\alpha}_c\mid\theta_r)$ 同时提供与 $\boldsymbol{\alpha}_c$ 有关的信息,而 θ_r 的信息则由概率密度函数 $P(\theta_r)$ 提供. 因此,联合后验分布 $\pi_i^{t-1}(\boldsymbol{\alpha}_c,\theta_r)$ 同时包含了 $\boldsymbol{\alpha}_c$ 和 θ_r 的信息,且 $\boldsymbol{\alpha}_c$ 的信息比 θ_r 的信息更丰富.

设 $P(X_{ij}=x\mid\boldsymbol{\alpha}_c,\theta_r)$ 为被试 i 在项目 j 上的作

答概率,任意 2 个作答概率之间的差异可用 KL 信息[19] 度量. 在 CD-CAT 中,KL 类的选题方法均是基于KL信息开发而得的,其中PWKL选题方法因其优越的选题性能和简洁的计算方式而被广泛使用. 因此,本文基于PWKL的选题思想,结合高阶模型的特性,拟开发适用于Higher-order CD-CAT、同时兼顾 α 和 θ 的 HO-PWKL 选题方法,其计算方法为

$$H_{\text{O-PWKL}_{ij}}=\sum_{r=1}^{R}\sum_{c=1}^{2^K}\Big(\sum_{x=0}^{1}P(X_{ij}=x\mid\hat{\boldsymbol{\alpha}},\hat{\theta})\log\big(P(X_{ij}=x\mid\hat{\boldsymbol{\alpha}},\hat{\theta})/P(X_{ij}=x\mid\boldsymbol{\alpha}_c,\theta_r)\big)\pi_i^{t-1}(\boldsymbol{\alpha}_c,\theta_r)\Big)\propto\sum_{r=1}^{R}\sum_{c=1}^{2^K}\Big(\sum_{x=0}^{1}P(X_{ij}=x\mid\hat{\boldsymbol{\alpha}})\log\big(P(X_{ij}=x\mid\hat{\boldsymbol{\alpha}})/P(X_{ij}=x\mid\boldsymbol{\alpha}_c)\big)\pi_i^{t-1}(\boldsymbol{\alpha}_c,\theta_r)\Big).$$

HO-PWKL 选题方法的目标函数为

$$O_{\text{bjective}}=\arg\max_{j\in B_i^{t-1}}H_{\text{O-PWKL}_{ij}},$$

其中 B_i^{t-1} 表示被试 i 在作答 $t-1$ 题后的剩余题库,从剩余题库中选取 $H_{\text{O-PWKL}_{ij}}$ 最大的项目 j 作为下一试题.

3　模拟实验

为了检验 HO-PWKL 与传统选题方法在 Higher-order CD-CAT 中的选题性能,本文开展了 2 项 Monte Carlo 模拟实验.

3.1　实验条件

实验 1 和实验 2 均包括 3 个自变量:(i) 测量属性个数(分别包含 5 个属性和 8 个属性);(ii) 测验长度(短测验和长测验),由于测量的属性个数越多,需要的项目就越多,因此 5 个属性的短测验和长测验分别设为 10 题和 20 题,8 个属性的短测验和长测验分别设为 15 题和 30 题;(iii) 选题方法,本文采用的选题方法均为 KL 类的方法,包括 KL、HKL、PWKL 和 MPWKL 等 4 个传统选题方法以及新开发的 HO-PWKL 选题方法. 由于 MPWKL 方法比较耗时,所以本文将采用 Zheng Chanjin 等[20] 提出的预先计算策略来提升 MPWKL 方法的运行速度.

3.2　被试参数和项目参数的模拟

所需 1 000 个被试从能力 θ 服从标准正态分布 $N(0,1)$ 的集合中随机抽取,每个被试的知识状态 α 则由 HO-CDMs 生成. 具体来说,通过式(1) 计算能力为 θ_i 的被试掌握每个属性的概率 $P(\alpha_k=1\mid\theta_i)$,然后将此概率和从均匀分布 $U(0,1)$ 中抽取的随机数 u 进行比较,若 $P(\alpha_k=1\mid\theta_i)\geqslant u$,则该被试掌握此属性,否则该被试没有掌握此属性. 以此类推,最终模拟出被试的真实知识状态. 在高阶参数方面,本文采用在已有研究中的参数设定[21-22],所有属性的斜率参数均固定为 $\lambda_1=1.5$;至于截距参数,当属性个数为 5 时 $\lambda_{0k}=(-1.0,-0.5,0,0.5,1.0)$,当属性个数为 8 时 $\lambda_{0k}=(-1.000,-0.715,-0.430,-0.145,0.145,0.430,0.715,1.000)$.

本文的题库容量设为 500 题,每个项目随机测量 1 ~ 3 个属性. 根据 Ma Wenchao 等[23] 的建议,从均匀分布 $U(0.1,0.4)$ 中为每个项目随机生成一个猜测参数 $P(0)$ 和一个失误参数 $1-P(1)$,$P(0)$ 表示在被试没有掌握项目测量的任何属性时猜对的概率,$P(1)$ 表示在被试掌握了项目测量的所有属性时正确作答的概率. 其他知识状态的答对概率从 $[P(0),P(1)]$ 中随机生成,并满足单调性约束.

给定被试真实知识状态、$\boldsymbol{Q}$ 矩阵和项目参数,通过 R 软件中 GDINA 包的 simGDINA 函数[23],为 5 个属性和 8 个属性的测试分别模拟一个 1 000 × 500 的完全作答矩阵. 为消除随机误差,每种模拟实验重复 30 次,计算 30 次实验的均值作为最终的实验结果.

3.3　评价指标

3.3.1　*模式判准率*　模式判准率是评价知识状态分类精度的指标,其值越大则精度越高,计算方法为

$$P_{\text{CCR}}=\sum_{i=1}^{N}I(\hat{\boldsymbol{\alpha}}_i=\boldsymbol{\alpha}_i)/N,$$

其中 $\hat{\boldsymbol{\alpha}}_i$ 为被试 i 知识状态的估计值,而 $\boldsymbol{\alpha}_i$ 表示其真实的知识状态;$I(\cdot)$ 为指向函数,当 $\hat{\boldsymbol{\alpha}}_i=\boldsymbol{\alpha}_i$ 时,计数为 1,否则计数为 0;N 为被试人数.

3.3.2　*能力估计精度*　均方差为能力估计精度的指标,其值越小则精度越高,其计算方法为

$$M_{\text{SE}}=\sum_{i=1}^{N}(\hat{\theta}_i-\theta_i)^2/N,$$

其中 $\hat{\theta}_i$ 为被试 i 能力值的估计值,θ_i 为其真实能力.

3.3.3　*题库使用均匀性*　本文将采用 χ^2 统计量和测验重叠率(test overlap ration,T_{OR}) 来衡量题库的安全性,χ^2 统计量的计算方法为

$$\chi^2=\sum_{j=1}^{J}(e_{r_j}-\bar{e}_{r_j})^2/\bar{e}_{r_j},$$

其中 J 为题库的总项目数;e_{r_j} 为项目 j 的实际曝光率,即项目 j 被使用的次数除以总人数;$\bar{e}_{r_j}$ 为项目 j 的期望曝光率,即 $\bar{e}_{r_j}=L/J$,L 为测验长度. χ^2 统计量越小表明项目的调用越均匀,题库也就越安全.

T_{OR} 反映不同被试调用相同项目的重叠情况，T_{OR} 值越高则题库越不安全. T_{OR} 的计算方法为

$$T_{OR} = N\sum_{j=1}^{J} e_{r_j}^2/((N-1)L) - 1/(N-1).$$

3.3.4 选题用时 选题方法的运行速度用选择每个项目的平均用时来表示，其计算方法为

$$T_C = \sum_{i=1}^{N} \overline{T}_i/N,$$

其中 $\overline{T}_i$ 为被试 i 选择每个项目的平均用时(单位为 ms). T_C 越小则选题速度越快.

4 实验结果

4.1 实验 1:在简化模型下各选题方法的比较

表 1 为当采用简化模型 HO-ACDM 时，各选题方法在不同条件下的被试参数估计精度.

表 1 在 HO-ACDM 模型下各选题方法的被试参数估计精度

选题方法	模式判准率(P_{CCR})				均方差(M_{SE})			
	5 个属性		8 个属性		5 个属性		8 个属性	
	短测验(10)	长测验(20)	短测验(15)	长测验(30)	短测验(10)	长测验(20)	短测验(15)	长测验(30)
KL	0.477	0.682	0.202	0.384	0.372	0.299	0.299	0.231
HKL	0.644	0.892	0.421	0.747	0.343	0.265	0.267	0.198
PWKL	0.645	0.892	0.422	0.747	0.344	0.265	0.266	0.198
MPWKL	0.650	0.895	0.428	0.751	0.342	0.264	0.262	0.197
HO-PWKL	0.762	0.926	0.613	0.836	0.312	0.254	0.233	0.183

在模式判准率(P_{CCR})方面，当测量 5 个属性时，新方法 HO-PWKL 在不同长度测验中的 P_{CCR} 都高于传统选题方法，这种优势在短测验中尤为明显. 如在短测验中，HO-PWKL 的 P_{CCR} 比精度最高的传统选题方法还要高 11.2%，而在长测验中，这种差距缩小到 3.1%. 8 个属性的结果与 5 个属性的结果相似，HO-PWKL 的 P_{CCR} 仍是最高的. 不同的是，当属性个数增加时，HO-PWKL 在 P_{CCR} 上的优势更明显. 以短测验为例，当测量 5 个属性时，HO-PWKL 与传统选题方法在 P_{CCR} 上的最小差距仅为 11.2%，而当测量 8 个属性时，这种差距增加到 18.5%. 这表明：测量的属性个数越多，HO-PWKL 在模式判准率上的优势越大.

能力的估计结果与知识状态相似，在不同条件下，HO-PWKL 的均方差(M_{SE})总是最低(即能力估计精度最高)，这种优势在短测验中也比较明显. 与 P_{CCR} 不同的是，HO-PWKL和传统选题方法在 M_{SE} 上的差异并不受属性个数的影响. 此外，随着测量属性个数的增加，所有方法的 M_{SE} 均显著降低，这是因为潜在二分属性在高阶能力估计中的作用类似二分项目，若测量的属性个数越多则高阶能力的估计精度越高[1]. 在传统选题方法方面，KL 的 2 个测量精度总是最差，HKL、PWKL 和 MPWKL 的测量精度均明显高于 KL 方法. 另外，MPWKL 的测量精度与 HKL 和 PWKL 相近，这和已有的结果不太一样，其原因可能是采用的模型不同.

表 2 为各选题方法在长测验中的题库使用均匀性指标. 由表 2 可知：HO-PWKL 的曝光率和测验重叠率总是最低. 这说明新方法的题库安全性比传统方法更高，而且这种优势不受测量属性个数的影响. 在传统选题方法中，HKL 和 PWKL 具有最低且相似的指标，MPWKL 的指标较高，KL 的 2 个指标最高，这表明 KL 的题库安全性最差.

表 2 在 HO-ACDM 模型下各选题方法的题库安全性

属性个数	曝光指标	选题方法				
		KL	HKL	PWKL	MPWKL	HO-PWKL
5	χ^2	365.892	321.259	321.566	343.169	289.320
	T_{OR}	0.772	0.682	0.683	0.726	0.618
8	χ^2	338.956	304.336	304.846	320.834	254.946
	T_{OR}	0.738	0.668	0.669	0.701	0.569

实验 1 的结果表明：在采用简化模型 HO-ACDM 的 Higher-order CD-CAT 中，新开发的 HO-PWKL 的测量精度和题库安全性均优于传统选题方法. 这说明新方法是在该 CAT 系统下较为理想的选题方法.

4.2　实验 2:在饱和模型下各选题方法的比较

表 3 为当采用饱和模型 HO-GDINA 时,各选题方法在不同条件下的被试参数估计精度,此结果和 HO-ACDM 的结果非常相似.

由表 3 可知:HO-PWKL 的测量精度依旧最高,其次是 MPWKL,再次是非常接近的 PWKL 和 HKL,最差的是 KL. 与传统选题方法相比,HO-PWKL 的 P_{CCR} 和 M_{SE} 在短测验中的优势同样突出;同时,若测量的属性个数越多,则 HO-PWKL 在 P_{CCR} 上的优势越明显. 与 HO-ACDM 结果不同的是,当采用 HO-GDINA 模型时,MPWKL 的 P_{CCR} 优于 PWKL 和 HKL 的,这和前人的研究基本一致,这说明模型确实会影响选题方法的选题性能.

表 3　在 HO-GDINA 模型下各选题方法的被试参数估计精度

选题方法	模式判准率(P_{CCR})				均方差(M_{SE})			
	5 个属性		8 个属性		5 个属性		8 个属性	
	短测验(10)	长测验(20)	短测验(15)	长测验(30)	短测验(10)	长测验(20)	短测验(15)	长测验(30)
KL	0.440	0.667	0.197	0.385	0.375	0.297	0.288	0.228
HKL	0.636	0.888	0.428	0.753	0.342	0.265	0.264	0.198
PWKL	0.635	0.887	0.431	0.757	0.344	0.265	0.263	0.197
MPWKL	0.652	0.898	0.455	0.769	0.339	0.260	0.257	0.194
HO-PWKL	0.741	0.921	0.592	0.838	0.316	0.254	0.238	0.187

表 4 为当采用饱和模型 HO-GDINA 时各选题方法在长测验中的题库使用均匀性指标. HO-PWKL 的曝光率和测验重叠率总是最小,这说明 HO-PWKL 在题库利用上比传统方法更均匀,即题库安全性更高.

表 4　在 HO-GDINA 模型下各选题方法的题库安全性

属性个数	曝光指标	选题方法				
		KL	HKL	PWKL	MPWKL	HO-PWKL
5	x^2	267.681	240.773	241.340	251.676	224.774
	T_{OR}	0.575	0.521	0.522	0.543	0.489
8	x^2	222.680	195.983	196.366	205.694	177.479
	T_{OR}	0.505	0.451	0.452	0.471	0.414

实验 2 的结果表明:在采用饱和模型 HO-GDINA 的 Higher-order CD-CAT 中,HO-PWKL 的测量精度在不同条件下都优于传统选题方法,同时题库安全性也更高. 综合来看,不论是饱和模型,还是简化模型,HO-PWKL 的测量精度都高于传统选题方法的测量精度,这表明此方法的选题性能不易受模型影响,比较稳定可靠.

表 5 为各选题方法选择每个项目的平均用时. 在所有条件下,KL 的选题用时均最少,其次是 PWKL 和 HKL,接着是新开发的 HO-PWKL,用时最多的是 MPWKL. 此结果和预期的结果一致,即方法越复杂,选题用时越多. 此外,选题方法的选题用时还与测量的属性个数及题库容量有关. 在固定题库容量情况下,属性个数越多,KL 类选题方法的选题用时越长,其中 MPWKL 的选题用时更是呈指数级增长[24]. 虽然本文采用了预先计算策略[20]提升了 MPWKL 的计算速度,但当测量的属性比较多时,MPWKL 的运算量仍旧非常大,因此选题用时最多.

表 5　各选题方法在每题上的平均用时　　ms

模型	属性个数	选题方法				
		KL	HKL	PWKL	MPWKL	HO-PWKL
HO-ACDM	5	9.576	9.743	9.620	17.428	14.962
	8	72.362	74.227	74.201	380.197	111.460
HO-GDINA	5	9.213	10.187	10.083	18.260	15.373
	8	86.793	87.656	87.568	389.299	123.989

5　总结与讨论

本文基于 PWKL 选题思路,结合高阶模型的特性,开发适用于 Higher-order CD-CAT、兼顾能力与知识状态的 HO-PWKL 选题方法. 模拟实验的结果表明:在不同条件下,HO-PWKL 的测量精度和题库安全性总是高于传统选题方法,尤其是在短测验或测

量多个属性时. 此外,新方法的优势并不受模型的影响,这表明新方法稳定可靠. 在选题用时方面,虽然 HO-PWKL 的用时比 KL、HKL 和 PWKL 更多,但即使在 8 个属性的测验中,HO-PWKL 选择每个项目的平均用时也低于 125 ms,这符合 CAT 的速度要求. 综合而言,本文开发的 HO-PWKL 基本可行,可以满足在 Higher-order CD-CAT 中兼顾能力和知识状态的选题要求,弥补了传统选题方法的不足.

虽然 HO-PWKL 的选题也兼顾了能力和知识状态,但它明显不同于传统双目标选题方法. 具体而言,由于在 Higher-order CD-CAT 中的项目不直接提供能力信息[1,8],因此 HO-PWKL 只考虑由先验和项目提供的知识状态信息以及能力的先验信息;而传统的双目标选题方法,同时考虑了由先验与项目提供的 2 个潜在变量信息. 总之,双目标 CD-CAT 使用的模型不同(单一模型或分离建模),选题方法的构造也不同[7].

与传统单目标选题方法相比,HO-PWKL 的测量精度更高的原因有 2 个:(i)HO-PWKL 在选题过程中考虑了能力信息;(ii)在 HO-PWKL 下的联合后验概率可能为知识状态提供了额外的信息,即知识状态的条件概率. 在题库安全性方面,本文推测被试划分的组数越多,项目的曝光率和测验重叠率就越低. 因此,与传统单目标选题方法相比,将被试划分为 $2^K \times R$ 个组的 HO-PWKL 的题库使用均匀性指标更低. 在选题用时方面,由于 HO-PWKL 要计算 R 次 PWKL 信息量,因此用时多于 KL、HKL 和 PWKL. 但与 MPWKL 方法相比,随着属性个数增多,HO-PWKL 方法仅增加 PWKL 信息量的运算,而在 MPWKL 中 $2^K \times 2^K$ 的 $\boldsymbol{D}$ 矩阵则呈指数级增长,其运算量也呈指数级增加,故而 HO-PWKL 的选题用时少于 MPWKL.

虽然本文开发的新选题方法具有比较理想的测量精度、题库安全性和运算速度,但未来研究还需在以下几方面进一步验证和拓展:首先,本文采用的高阶模型参数和项目参数均假定为真实值,然而在实际中的参数标定不可避免地存在标定误差,其中高阶模型参数的标定误差可能更大[25-26]. 因此,还需要进一步研究标定误差对 HO-PWKL 选题性能的影响. 其次,由于多维度测量和多级评分项目已广泛应用于心理与教育测验中[27-30],所以,为了推动 Higher-order CD-CAT 更好地服务实际,有必要开发适用于多维多级的 Higher-order CD-CAT 选题方法,或者将单维 2 级的 HO-PWKL 拓展到多维多级. 最后,本文是在定长 Higher-order CD-CAT 中检验 HO-PWKL 的选题性能,然而变长 Higher-order CD-CAT 可能更符合实际需要,即通过不同长度的测验获取相同的测量精度. 因此,有必要进一步探究 HO-PWKL 在变长 Higher-order CD-CAT 下的测验效率.

6 参考文献

[1] HSU C L, WANG Wenchung. Variable-length computerized adaptive testing using the higher order DINA model [J]. Journal of Educational Measurement, 2015, 52(2): 125-143.

[2] WANG Chun, ZHENG Chanjin, CHANG Huahua. An enhanced approach to combine item response theory with cognitive diagnosis in adaptive testing [J]. Journal of Educational Measurement, 2014, 51(4): 358-380.

[3] GREENO J G. Trends in the theory of knowledge for problem solving [EB/OL]. [2021-12-18]. https://www.researchgate.net/publication/244520649_Trends_In_the_theory_of_knowledge_for_problem_solving.

[4] TEMPLIN J L, HENSON R A. Measurement of psychological disorders using cognitive diagnosis models [J]. Psychological, Methods, 2006, 11(3): 287-305.

[5] CHANG Huahua. Making computerized adaptive testing diagnostic tools for schools [EB/OL]. [2021-12-18]. https://www.mendeley.com/catalogue/f222f893-6e44-3fcf-b3ab-661f2fce7047/.

[6] CHENG Ying. The dual information method for item selection in cognitive diagnostic computerized adaptive testing [D]. Urbana-Champaign: University of Illinois at Urbana-Champaign, 2007.

[7] 罗芬, 王晓庆, 蔡艳, 等. 基于基尼指数的双目标 CD-CAT 选题方法 [J]. 心理学报, 2020, 52(12): 1452-1465.

[8] HUANG Hungyu. Utilizing response times in cognitive diagnostic computerized adaptive testing under the higher-order deterministic input, noisy ´and´ gate model [J]. British Journal of Mathematical and Statistical Psychology, 2020, 73(1): 109-141.

[9] WANG Chun, CHANG Huahua, DOUGLAS J. Combining CAT with cognitive diagnosis: a weighted item selection approach [J]. Behavior Research Methods, 2012, 44: 95-109.

[10] de la TORRE J, DOUGLAS J A. Higher-order latent trait models for cognitive diagnosis [J]. Psychometrika, 2004, 69: 333-353.

[11] DAI Buyun, ZHANG Minqiang, LI Guangming. Exploration of item selection in dual-purpose cognitive diagnostic computerized adaptive testing: based on the RRUM [J]. Ap-

plied Psychological Measurement,2016,40(8):625-640.
[12] KANG HYEON-AH,ZHANG Susu,CHANG Huahua. Dual-objective item selection criteria in cognitive diagnostic computerized adaptive testing [J]. Journal of Educational Measurement,2017,54(2):165-183.
[13] MCGLOHEN M,CHANG Huahua. Combining computer adaptive testing technology with cognitively diagnostic assessment [J]. Behavior Research Methods,2008,40(3):808-821.
[14] ZHENG Chanjin,HE Guanrui,GAO Chunlei. The information product methods:a unified approach to dual-purpose computerized adaptive testing [J]. Applied Psychological Measurement,2018,42(4):321-324.
[15] 涂冬波,蔡艳,戴海琦. 几种常用非补偿型认知诊断模型的比较与选用:基于属性层级关系的考量 [J]. 心理学报,2013,45(2):243-252.
[16] TATSUOKA K K. A probabilistic model for diagnosing misconceptions by the pattern classification approach [J]. Journal of Educational Statistics,1985,10(1):55-73.
[17] de la TORRE J. The generalized DINA model framework [J]. Psychometrika,2011,76:179-199.
[18] JUNKER B W,SIJTSMA K. Cognitive assessment models with few assumptions,and connections with nonparametric item response theory [J]. Applied Psychological Measurement,2001,25(3):258-272.
[19] COVER T M,THOMAS J A. Elements of information theory [M]. New York:Wiley,1991.
[20] ZHENG Chanjin,CHANG Huahua. High-efficiency response distribution-based item selection algorithms for short-length cognitive diagnostic computerized adaptive testing [J]. Applied Psychological Measurement,2016,40(8):608-624.
[21] de la TORRE J,LEE Y S. A note on the invariance of the DINA model parameters [J]. Journal of Educational Measurement,2010,47(1):115-127.
[22] de la TORRE J,HONG Y,DENG W. Factors affecting the item parameter estimation and classification accuracy of the DINA model [J]. Journal of Educational Measurement,2010,47(2):227-249.
[23] MA Wenchao,de la TORRE J. GDINA:an R package for cognitive diagnosis modeling [J]. Journal of Statistical Software,2020,93(14):1-26.
[24] KAPLAN M,de la TORRE J,BARRADA J R. New item selection methods for cognitive diagnosis computerized adaptive testing [J]. Applied Psychological Measurement,2015,39(3):167-188.
[25] de la TORRE J. DINA model and parameter estimation:a didactic [J]. Journal of Educational and Behavioral Statistics,2009,34(1):115-130.
[26] PATTON J M,CHENG Ying,YUAN Kehai,et al. The influence of item calibration error on variable-length computerized adaptive testing [J]. Applied Psychological Measurement,2013,37(1):24-40.
[27] 蔡艳,苗莹,涂冬波. 多级评分的认知诊断计算机化适应测验 [J]. 心理学报,2016,48(10):1338-1346.
[28] 韩雨婷,高旭亮,汪大勋,等. 多级评分项目的多维 CAT 选题方法开发 [J]. 心理科学,2018,41(6):1500-1507.
[29] 康春花,辛涛. 测验理论的新发展:多维项目反应理论 [J]. 心理科学进展,2010,18(3):530-536.
[30] CHENG Ying. When cognitive diagnosis meets computerized adaptive testing:CD-CAT [J]. Psychometrika,2009,74(4):619-632.

The Item Selection Method Considering the Ability and Cognitive Profile in Higher-Order CD-CAT

XI Chongqin,TU Dongbo,CAI Yan*
(College of Psychology,Jiangxi Normal University,Nanchang Jiangxi 330022,China)

Abstract:Current item selection methods used in the CAT system select items adaptively according to only the attribute profile,which may lead to low precision with respect to ability. In light of this,under the item selection method of the higher-order CD-CAT and PWKL,the new item selection method that simultaneously considers the ability and attribute profile information is proposed for the higher-order CD-CAT in the paper. The results from the simulation study indicate that the new method proposed in this study always outperforms the existing methods in terms of measurement accuracy of the ability and cognitive profile,and the proposed method has also a significant advantage over the traditional methods in item pool security.

Key words:dual-objective CD-CAT;higher-order model;higher-order CD-CAT;item selection method

(责任编辑:冉小晓)